深層学習と言語学の接点
―中国語助動詞“会”と“要”を例に―

石田　智裕

The connection between deep learning and linguistics
—The case of the Chinese auxiliary verbs HUI and YAO—

ISHIDA Tomohiro

Abstract

This paper explores how deep learning can be utilized in the field of linguistics. Utilizing AI technologies is a huge challenge in modern society. The Japanese government considers it necessary to increase the number of AI engineers and data scientists. Most language research has been conducted by NLP researchers; in contrast, linguists are currently seeking ways of using AI technologies such as deep learning for linguistic research.

In this paper, the author makes use of word2vec (a kind of deep learning method) to attempt to discriminate two Chinese auxiliary verbs HUI and YAO which have been pointed out by Chinese linguists as having similar functions.

In section one, the author explains how other fields are using AI for their own research. A technical description of word2vec and the method of adopting it for linguistic research is also introduced.

In section two, previous studies about word2vec and the Chinese auxiliary verbs are introduced.

In section three, the author implements addition and subtraction of HUI and YAO by means of word2vec. By vectorizing natural language words, the difference between the two auxiliary verbs is calculated. By comparing the outcome of the mathematical operations and traditional Chinese linguistic research, the author shows that the machine learning operation and the fruits of previous Chinese linguistic studies reach partially similar results.

In section four, the author summarizes the implications of this study and explains how for linguistic research, it is necessary to control morphological analyses to obtain more fruitful outcomes.

目次

1. 研究背景

1.1. AI 技術の発展の概観

第 3 次 AI ブームと呼ばれる AI 化の波が押し寄せている。コンピュータの処理能力の向上によって、AI 技術の普及は広まっている。AI 技術に由来する一部のツールは、家庭向けの PC やスマートフォンでも十分に動かすことができるなど、その導入コストとハードルは極めて低くなっていると言える。画像認識や音声認識ではすでにハイレベルな解析技術が達成されている[i]。

自然言語の解析においては、自然言語処理という学問領域で深層学習の活用が進んでいる。例えば 2016 年に改善された google 翻訳においては、ニューラルネットワークを用いた機械学習が使用されている[ii]。

一方で、言語学・特に文法研究においては、AI 技術の利活用はまだ模索段階である。仮に AI 技術を導入するとしても、どのような研究課題に対して、どのように AI を使用することが適切なのか、その費用対効果は見合っているのかなど、議論しなければいけない問題は多岐にわたっている。

1.2. 自然言語を数字に変換する：word2vec

word2vec（word to vector: 単語からベクトルへの意）とは、Tomas Mikolov らの研究チームの研究によって確立された自然言語処理の手法である（Mikolov et al. 2013）。単語の分布を、その前後の語との共起関係を利用して、ニューラルネットワークを通じて解析、教師なしで学習することができる。深層学習の結果として、単語はベクトルに変換される。そのため、「word to vector」の名がつけられている。

word2vec は、語の分布に着目している。文中で交換可能な語は同じような意味や働きを持つ語であるという観点から学習していくのである。

word2vec は、単語をベクトルに変換する。この過程を、単語の埋め込み（word embedding）や、単語の分散表現（distributed representation）と呼ぶ。

学習の方法には、CBOW（Continuous Bag-of-Words Model）と Skip-gram（Continuous Skip-gram Model）の二種類がある（Mikolov et al. 2013 p. 4-5）。

CBOW は、ターゲット語 w(t) となる語の前文脈

w(t-2)[iii], w(t-1) と、後文脈 w(t+1), w(t+2) をニューラルネットワークの入力データとして、そこからターゲット語を推測する。その後、実際のターゲット語を参照し、推測したものと実際の語のベクトルの差異を見ることによって、どの程度推論が当たっているのかを確認していくという手順で学習を進めていく。つまり、前後文脈からターゲット語を推測していく学習方法である。

Skip-gram では、ターゲット語 w(t) がまず与えられ、ニューラルネットワークの入力データになる。その上で、前文脈 w(t-2) , w(t-1) と、後文脈 w(t+1), w(t+2) を推測し、学習する。つまり、ターゲット語から前後にどんな語が出現しうるかを推測していく学習方法である。

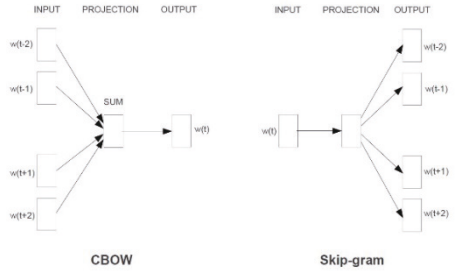

図 1（Mikolov et al. 2013 p. 5）

word2vec の特徴として、単語をベクトルという実数値に変換することが挙げられる。数値であるから、数学的な操作が可能になる。例えば、単語の加減算が可能であることが有名である。この「ベクトル」とは、言語学の観念で言えば、意味特徴と言ってもよい。意味特徴をコロケーションを頼りに機械学習し、言語ではなく数字で表したものがベクトルである。

具体例として、Mikolov et al. (2013) では、*"Paris – France + Italy = Rome"* が挙げられている。この操作は、「パリ」の持つベクトルから「フランス」の持つベクトルを引き、「イタリア」の持つベクトルを足すという操作によって、「ローマ」を得ている。言語学の観念で言えば、「パリ」の持つ意味特徴から「フランス」を引くことによって、まず「首都」という意味特徴を残している。その後で、「イタリア」の持つ意味特徴を加算することで、「首都」＋「イタリア」の意味特徴に最も近い語である「ローマ」を産出しているのである。（p. 9）

同論文で公開された *"king – man + woman = queen"* も著名な例である（p. 2）。「（男性の）王様」のベクトル（≒意味特徴）から、「男性」のベクトルを引くことで、「王権を持っている者」のベクトルが残ることになる。そこに「女性」を足せば、「女王」が産出されるのである。

word2vec における単語の加減算とは、単語のベクトルの加減算であり、換言すれば意味特徴の加減算であると言える。

1.3. 深層学習は文法研究に寄与するか？

近年の「AI ブーム」によって、様々な分野で AI 技術を取り入れようという機運が高まっている。日本国としても、政府が AI 技術者・データサイエンティストの育成を推進するなど、AI の利活用を進める方針を打ち出している[iv]。

しかしながら、闇雲に AI を利用したところで、研究課題が不適切であれば研究に付加価値を生じない。また、もし AI 技術の導入コストが極めて大きく、それに比べて研究に与える付加価値が小さいとすれば、AI を使用するメリットはないと言えるだろう。

言語学・特に文法研究の分野において AI 技術・深層学習を活用するためには、AI 技術が得意とし、ヒトの研究者が苦手とするのはどんな作業であるかを見定める必要がある。

AI 技術がヒトの研究者より優れている点としては、文章を処理する速度とその正確性が挙げられる。例えば word2vec を使用する場合、適切な言語データを収集し、モデル化しさえすれば、数百億語の文字列を容易に処理することができる。これは、人間の研究者では物理的に不可能なことである。また、プログラムに瑕疵がない限り、文字の見落としをすることもない。すなわち、ヒューマンエラーが起こりづらい点も AI 技術の優位点である。

コーパス言語学の隆盛など、言語資料の電子化の流れは既に存在している。ただし、これまでのコーパス研究はあくまで人間が言語を分析するもので、コーパスは言語資料の保管と検索のために存在していた。深層学習によって膨大な言語資料を短時間で分析できるAI技術を使用することは、大幅な時間の短縮とヒューマンエラーの削減というメリットをもたらすだろう。

ただし、これらは適切なデータを学習させ、モデル化することができるという前提のもとではじめて達成できることである。そのためには人間の研究者が適切にデータを選定、成形する必要があることは言うまでもない。

「言語」をAI技術で分析するという研究課題は、従来、自然言語処理（Natural Language Processing）という領域で研究されてきた。自然言語処理では、言語現象の中でも、内容語を中心として研究を積み重ねてきた（王棟ら 2019 p. 32）。そのため、名詞や形容詞の弁別などは一定の研究が積み重ねられている。その一方で、機能語（助詞など）の研究はあまり事例がない。

言語学・特に文法論では、機能語の研究は非常に大きな研究テーマである。大量の言語データを漏れなく短時間で処理する深層学習は、従来のコーパス研究より速く、正確に機能語の使用実態を映し出せる可能性がある。本研究では、中国語の助動詞の類義語弁別において、word2vecを使用し、分析を試みる。

1.4. 中国語の類義語“会”vs“要”

中国語の助動詞（能愿动词）“会（huì）”と“要（yào）”は、どちらも多義語である。“会”は能力を、“要”は意志や義務を表す語であり、異なった意味特徴を持つ。その一方、ある事象の実現する可能性を判断するような近似のモダリティ用法を持っており、同じ文内でどちらでも使用できる場合も多く、類義語であるとも言える。郭继懋 , 郑天刚编（2002）や王牧（2018）など、これらの語の用法を比較する研究も存在しており、中国語助動詞研究の論点の一つとなっている。

本研究では、この二種類の助動詞の意味の弁別に、どの程度word2vecが使用できるのかを実験を通して確かめる。

1.5. 研究課題

RQ1 助動詞“会”と“要”の弁別にword2vecは役立つか。

RQ2 word2vecを用いた類義語研究に、モデルのデータサイズの多寡はどの程度影響を与えるか。

2. 先行研究

2.1. word2vec

前述の通り、語同士の加減算はword2vecの基礎的な機能であり、word2vecの黎明期であるMikolov et al. (2013)の段階ですでに達成されている。一方、文法研究の文脈においてword2vecを使用した例としては、内田（2018）, 内田（2019）などが存在しているが、言語学の文法研究のための利活用を目的とした研究は盛んではない。

2.2. 助動詞“会”と“要”

“会”と“要”は、いずれも現代中国語の助動詞（能愿动词）である。いずれも未実現事象を目的語として使用されることが多く、特に事象の実現可能性を判断する文においては交換可能な場合がある。

2.2.1. “会”は能力・可能性

“会”は、主に能力·可能性を表すとされている。《现代汉语八百词（增订版）》では、“懂得怎样做或有能力做某事。”, “善于做某事。前面常加“很、真、最”等。”, “有可能。 通常表示将来的可能性, 但也可以表示过去和现在的。”の三用法が挙げられている。これは、「できる」「長けている」「可能性がある」に“会”を三分する考え方であり、《现代汉语词典》もこの分類を取っている。“会”の機能について、能力を表すものにつ

いては“能”,“可以”などとの区別から論じられる傾向が強い。

可能性を表す“会”については、主に推測・可能性判断のモダリティという観点から分析されてきた。“会”は、目的語になる事象が実現する可能性を話し手が判断するという意味を持つ。王其莉（2015）では、“会”の背景には必ず条件性があり、条件を受けての発話であることが“会”の中核的な意味であると述べられている（p. 141）。王牧（2018）では、“会”を用いた可能性判断は何らかの情報源から行われる論理的なものであり、直感的なものではありにくいとされている（p. 34）。安本（2019）では、“会”の機能について、「時間経過によって変わりにくい、恒常的な性質·能力」を表すとしている。これは「能力」や「可能性」に限らない、“会”の持つ意味特性の 1 つであるとしている。

2.2.2. “要”は意志・義務・瞬間・可能性

“要”もまた、使用頻度の高い助動詞の 1 つである。《现代汉语八百词（增订版）》では、以下の五つの意味項目に分類されている。“表示做某事的意志”,“须要应该”,“表示可能”,“将要”,“表示估计 , 用于比较句”、つまり「意志性」「義務性」「可能性」「短時間」「比較文での推測」の 5 種類である。“要”は主に話し手の意志や義務といった deontic modality を担当している点で“会”と異なっている。一方、可能性判断という部分では epistemic modality を表すことができ、この部分では“会”と類似していると言える。

2.2.3. “会”と“要”の接点：可能性判断

主に未実現事象を表す際に使われる点で類似し、交換可能である場合も多い。

(1) a 看样子，明天会下雨。（朱 1982 p. 63）
　　b 看样子 , 明天要下雨[v]。（作例）
　　この様子では、明日は雨が降りそうだ。

（1）では、どちらも未実現の事象“下雨”を目的語にとっており、その事象がまだ発生していないことを表わしていると言える。一方、“会”と“要”が入れ替えにくい例も存在している。

(2) a　你这样下去，身体会吃不消。（作例）
　　b　你这样下去，身体要吃不消。（作例）
　　このままでは体を壊すよ。

（2）では、「体を壊す」という事象はまだ未実現の事象である。つまり、“会”も“要”もどちらを使っても一見すると問題がないかのように見える。しかしながら、2b は違和感の残る文であるとする母語話者が、5 人中 4 人を数えた。それは、“要”を使うことによって「話し手の意志性」が現れてしまうからである。つまり、「この話し手は相手に体を壊してほしいのだ」というニュアンスが出てしまい、これは単純な可能性判断ではありにくい。“要”の持つ「可能性判断」以外の意味項目の影響を受けていると言えるだろう。

王牧（2018）では“会”と“要”の差異について、どちらも推測に使用できるが、“会”を用いた推測は論理的で、非現実の色彩があるが、“要”を用いた推測はより直感的であるという点を挙げている（p. 25-32）。また、“要”の主語は特定の主語を取りやすいが、“会”は不定の主語を取りやすいという特徴も指摘されている（p. 24）。

“会”と“要”は「可能性判断」という点では重なる点もあるが、それぞれの持つ他の意味項目の影響があるため、両者は完全に同一ではないと言える。

3. word2vec を用いた“会”と“要”の加減算

3.1. 言語モデルと加減算の結果

本章では、word2vec を使用し、助動詞“会”と“要”の加減算を行う。実験の手順としては、3.2 において“会”＋“要”、3.3 において“会”-“要”、3.4 において“要”-“会”を行う。得られた単語リストを言語学·中国語学における先行研究と照合し、どの程度一致し

ているかを考察する。

実験に使用するモデルは、a. 苏（2017）で作成された約 650 億語のモデル、b. 王棟ら（2019）において作成された、約 22 億語のモデルの二種類を使用する。

表 1

モデル	苏(2017)
レジスター	論説文、ニュース、技術報告
述べ語数	650億
学習方法	Skip-gram, window size 10

表 2

モデル	王棟ら(2019)
レジスター	小説
述べ語数	22億語
学習方法	Skip-gram, window size 5

苏（2017）のモデルは、最大手の中国語 SNS アプリケーション“微信”に内蔵されたミニブログ機能である“公众号”から収集されている。レジスターは論説文・ニュース・化学技術文が中心となっている。

王棟ら（2019）のモデルは、現代語で書かれた小説で構成されており、特別な設備は用いず一般向けの PC で学習させたモデルである。こちらは、個人の研究者でも特別な設備投資なしで作成できる規模のモデルであり、この水準のモデルがどの程度まで研究に利用できるかをはかる意味で使用する。

また、異なるレジスターを用いることで、より実験の結果が中国語の実態を反映するようになることが期待される。

この実験の目的は、深層学習によって得られた言語のベクトルデータによる分析が、既存の言語学の研究とどの程度まで符合するかを観察することである。

換言すれば、人間の研究者により積み重ねられた研究成果を、特別な設備投資をせずに一般向けの PC で利用できる、無料公開されているモデルで、どこまで再現でき、どの部分では及ばないのかを確かめることにある。

既存のモデルを活用するのであれば word2vec は誰もが導入コストゼロで使用できる。もし苏（2017）のような既存の大型モデルが言語学の先行研究に符合するデータを産出できるなら、言語分析の新たなツールとして使用価値があると言えるだろう。

また、個人が特別な設備投資なしで作成した王棟ら（2019）のモデルが実用に足りるとすれば、word2vec の利活用のハードルはさらに下がり、各研究者が個人で研究用モデルを作成し、分析するビジョンが見えるようになるだろう。

また、仮に深層学習で得られたデータが一定のレベルで人間の分析に近づいているのであれば、今後の言語学研究においては、先に深層学習を用いて言語のデータを解析しておき、それをヒント・下敷きにして研究を深めていくという研究手法を取ることが可能になる。

3.2. “会”+“要”

“会”+“要”については、次頁表 3 のような結果が得られた。

苏（2017）については、接続詞・助動詞・副詞などが検出された。“会”の文脈でも“要”の文脈でも使用できるという語である。

［述詞性成分に前置される語やフレーズ］
“不会（できない , ありえない）”, “一定（必ず）”, “应该（…べきだ）”, “可能（かもしれない）”, “就要（”要“とほぼ同義）”, “需要（必要がある）”, “必须（…すべきだ）”, “还要（まだ必要がある）”, “不能（できない）”, “不要（してはいけない･なくてもよい）”, “一旦（いったん）”, “肯定（必ず）”

表 3

	会＋要	
	苏(2017)	王棟ら(2019)
1	不会	可以
2	所以	他们
3	一定	我们
4	应该	你
5	可能	是
6	就要	也
7	不然	就
8	需要	他
9	因为	不
10	如果	能
11	必须	了
12	还要	让
13	不能	自己
14	不要	想
15	一旦	我
16	那么	和
17	,	对
18	肯定	现在
19	的	你们
20	总之	人

［接続詞］

“不然（さもないと）”,“因为（…だから）”,“如果（もし）”,“那么（…であれば）”,“总之（とにかく）”

［その他］

“,[vi]”,“的（の：連体修飾のマーカー）”

全体に機能語がリストアップされていることがわかる。その中でも、“不会”と“不要”が挙げられていることについては、助動詞の否定形（“不”を付加した形）は肯定形と交換可能である場合が多いことを考えれば当然ともいえる。一方、中国語学においては“不”と“会”、“要”の合計二語と分析することが多く、これを一語と見做す形態素解析の方法は言語学的にはそぐわないと言えるかもしれない。

“一定”,“肯定”という高い確度での可能性判断を表すモダリティ副詞もリストには含まれている。“会”と“要”はいずれも可能性判断に使われることから、このようなモダリティ副詞と近い意味特徴を持っていることも頷けるだろう。しかしながら、“应该”のような、義務モダリティを表す語が“会”と“要”双方のベクトルに近似しているということは特筆に値するだろう。“要”には義務の用法があることが多く指摘されているが、“会”にはそのような指摘は見られないからである。この事実は、後述の“要”-“会”の結果とやや矛盾すると言える。

全体に述詞性成分に前置する語が多く含まれていることは、“会”及び“要”が助動詞であり、基本的には述詞性成分（主に動詞）を目的語に取るという性質を持っていることからの影響であろう。即ち、助動詞の統語的な性質は、深層学習によって十分に学習されていると言える。

接続詞に関しては、“那么”は主に条件文で使用され、条件文の後件に置かれて論理や時間の前後を表すことが多い。“会”も“要”も、論理的・時間的に後に発生する事象を目的語に取る傾向が強いため、どちらも“那么”と入れ替え可能な場合が多くあり、近いベクトルを持つことは首肯できる。しかしながら、その他の接続詞については“会”や“要”の持つ性質との関係は見出しづらく、議論が必要であると言える。

その他、論理関係を区切る符合である“,（逗号）”がリストアップされていることも興味深い。“会”や“要”が論理・時間の前後関係を表すということは、換言すれば論理や時間に区切りを作る機能を持つということである。これは、文に区切りを打つ“,”の機能と類似していると言うことができる。そのため、近いベクトルを持つことは不思議ではない。

構造助詞“的”については、“会”や“要”と入れ替え可能な文脈は一般的ではない。“的”の直後は基

本的に体詞性成分が現れるのに対して、助動詞である“会”や“要”は述詞性成分を目的語に取りやすいからである。そのため、統語的に見れば“会”や“要”と“的”が交換可能である文脈は多くないと推測できる。この“的”については更なる議論が必要である。

王棟ら（2019）のモデルでは、代名詞が非常に多いことが見て取れる。

［代名詞］
“他们（彼ら）”,“我们（私たち）”,“你（あなた）”,“自己（自分･自ら）”,“我（私）”,“你们（あなたたち）”,“人（人）”

［述詞性成分を目的語とする語］
“可以（できる･してもよい）”,“也（…も）”,“就（時間の短さ･前後関係を表す）”,“不（not）”,“想（…したい、思う）”

［助詞］
“了（完了・実現・変化などのマーカー）”

［動詞］
“是（英語の be に相当）”,“让（使役動詞）”

［接続詞］
“和（…と）”

［前置詞］
“对（…に対して、或いは「yes」）”

［その他］
“现在（今）”

“会”とも“要”とも近い意味特徴を持つかどうかと考えたとき、代名詞の場合は疑問が残る。品詞や性質がまったく違うからである。しかしながら、純粋に語順だけを見れば両者の分布は近い。

(3) 我会去你家。（作例）
あなたの家に行くよ。

(4) 我去你家。（作例）
あなたの家に行くよ。

“会”や“要”は、文の主語に近い位置に分布することが可能である。すると、上記の文で言えば“会”と“我”は近い位置に分布しており、語順だけで見れば類似表現であるともいえる。これも、“会”や“要”の統語的な性質を反映していると言えるだろう。

その他、助詞の“了”が出現している点は、中国語学の立場からすれば大きく議論の余地がある。第一に、未実現事象の実現可能性を表す“会”は、すでに実現したことを表すアスペクト助詞の“了 1”とは逆の性質を持ち、共起しづらいことが指摘されている（朱 1982 p. 63）。そのため、“了”が“会”の前後 5word の中に相当数出現すること自体に疑問の余地があるともいえる。一方で、能力の“会”であれば“了 1”との共起も可能である。また、“了”の方が語気助詞の“了 2”であれば“会”との共起も可能である。問題なのは、このデータからでは、果たしてこの“了”は“了 1”なのか“了 2”なのかはっきりしないことである。また、“会”も「能力」と「可能性」のどちらの用法の物なのかを区別せずに学習しているため、厳密な文法の議論には結び付けづらいだろう。この点については、データの前処理の段階で文法議論に適した処理を施す必要が出てくる。例えば、“了”であれば、少なくとも動詞の直後に分布し、句点の直前でなければ“了 1”のタグをつけ、句点の直前であれば“了 2”のタグをつける、動詞の直後の“了”で文が終わっていれば“了 1 + 2”のタグをつけるなどすれば、大きな手間はなくこれらを区別できるだろう。ただし、文法分析のための言語データの成形は決して単純ではないことは 4.2 で改めて記述する。

3.3. “会”-“要”

表 4

	会–要	
	苏(2017)	王棟ら(2019)
1	久而久之	可能
2	长此以往	或许
3	莫名	窝藏
4	长此下去	流落
5	消失	甚么样
6	相形见绌	不祥
7	意想不到	伤怀
8	会演	神规
9	可能	情分
10	淌眼泪	乐趣
11	讶异	善意
12	蒸干	凡间
13	没来由	预报
14	变得	探视
15	出现	多管闲事
16	遇热	造成
17	产生	孤单
18	心绪不宁	柔弱
19	暗淡无光	破戒
20	不经意间	重返

苏（2017）の“会”-“要”の内容を見てみると、意味特徴ごとに以下のように分けることができるだろう。

［+ 長時間］
“久而久之（時間の経つうちに）”,“长此以往（長い間続く）”,“长此下去（長い間続く）”

［+ 推測］
“可能（かもしれない）”

［- 意志］
“莫名(なんとなく)”,“消失(消える)”,“相形见绌(大きく劣る)”,“意想不到(思いがけない)”,“淌眼泪(涙を流す)”,“讶异(驚く)”,“蒸干(蒸発乾燥する)”,“没来由（なんとなく）”,“变得（変化する）”,“出现（出現する）”,“遇热（熱源に触れる）”,“产生（生まれる）”,“心绪不宁（心が乱れる）”,“暗淡无光（暗い・彩りがない）”,“不经意间（知らない間に）”

［その他］
“会演（合同公演をする）”

［+ 長時間］の三語が、“会”-“要”の上位に現れている。≪八百词≫によれば、“要”には“将要”の意味が含まれている。即ち、ある事象が短い時間のうちに発生することを表す際に“要”を使う傾向があると指摘されている。そのため、“会”から“要”を減算する際に、“会”の文脈にあって“要”の文脈にないものとして、［+ 長時間］の語が現れることは十分に考えられるだろう。

［+ 推測］の語は、苏（2017）のモデルでは“可能”のみである。“会”と“要”のどちらにも「可能性判断」のモダリティ機能があるとする先行研究に対して、“会”のベクトルのみ“可能”が近いというのはやや矛盾していると言える。

［- 意志］の語句については、いくつかの下位分類に分けることができるだろう。第一に、「思いがけない」というように、話し手の意志・思考が介在していないことを表す語である。“莫名”,“意想不到”,“没来由”,“不经意间”がこれに該当する。これらはいずれも、話し手の意志や思考と無関係になんらかの事象が存在する際に使用される語である。明確に、話し手の意志が含まれないと言えるだろう。

第二に、出現や消滅を表す動詞である。“消失”,“变

得”,“出现”,“产生”がこれに該当する。事象の出現・消滅を表すこれらの動詞は、動作主の意志性が顕著でない類の動詞であると言える。

第三に、意志的でない心理動詞や動詞フレーズである。“淌眼泪”,“讶异”,“心绪不宁”がこれに該当する。これらの動詞（フレーズ）は心理活動やそれに付帯する動作を表す。いずれも通常は動作主が意志的にこのような心理状態になるとは考えにくく、不如意の心理状態であると言える。

第四に、マイナスの状態を表す語である。“相形见绌”,“暗淡无光”がこれに該当する。この解釈はやや難しいが、このようなマイナスのニュアンスを持つ語は、動作主ないし被修飾語が望んでなるものとは言い難く、これも不如意の物であると考えられる。

［その他］には、“会演”を分類した。この語には、時間や意志性に関わるニュアンスがない。「合同公演であるため、動作主の意志ではなく合同公演者の意向に忖度せねばならない」というような形で無意志性が反映されている可能性も無きにしも非ずであるが、語の中に“会”の字が入っていることを考えると、“会演”を“会”と誤認した形態素解析レベルでの間違いである可能性が高いように思われる。

王棟ら（2019）モデルの“会”-“要”結果は、苏（2017）モデルを支持するものもあるが、そうでない部分も存在する。

［+ 推測］
“可能（かもしれない）”,“或许（かもしれない）”

［- 意志］
“流落（落ちぶれる）”,“不祥（不吉だ）”“伤怀（悲しむ）”,“造成（引き起こす）”,“孤单（孤独だ）”,“柔弱（軟弱だ）”

［その他］
“窝藏（かくまう）”,“甚么样（どのような）”,“神规（身分の一種）”,“情分（義理人情）”,“乐趣（愉しさ）”,“善意（善意・好意）”,“凡间（世間・人間社会）”,“预报（予報）”,“探视（お見舞い）”,“多管闲事（不要なところででしゃばる）”,“重返（復帰する）”,“破戒（戒律を破る・破戒）”

［+ 推測］については、苏（2017）モデルでもあった“可能”に加えて、“也许”が含まれている。どちらも推測のモダリティを持つ代表的な語であり、［+ 推測］であることは間違いない。“要”にも可能性判断の用法がありながら、なぜ“会”-“要”をするとこの種の語が残るのかについては次節で考察する。

［- 意志］の語については、貶義語が複数残った。動詞・形容詞に大別出来るが、いずれにせよ動作主の意志的な動作行為でないことは間違いないと言える。

上記のどちらにも分類しづらい［その他］の語も多い。“窝藏（かくまう）”,“探视（お見舞い）”,“多管闲事（不要なところででしゃばる）”などは動作主が意志的に行うことの多い動作と推測でき、［- 意志］とは言い難い。その他の語も、一見すると“会”や“要”の持つモダリティとどのような関係があるのか想像しづらい。

このような結果が得られた理由については、1 つにはデータの規模の問題があり得るが、3.5 にて詳説する。

3.4. “要”-“会”

“要”-“会”については、次頁表 5 のような結果が出ている。

“要”-“会”の結果得られた語については、まず苏（2017）のモデルから見てみると、他者に何らかの訓戒を与える際に使用されるような語句が非常に多いことが見て取れるだろう。大まかに分けるとすれば、［+ 義務］［+ 意志］のように区分できる。

［+ 義務］
“必须（しなければならない）”,“尤须（特に……すべきだ）”,“要端正（きちんとせねばならない）”,“要

表 5

	要–会	
	苏(2017)	王棟ら(2019)
1	分清主次	扎寨
2	必须	斩
3	牢记	华雄
4	摆正位置	郡
5	注意	华佗
6	尤须	老将
7	认清	孔融
8	突出重点	太守
9	做到	关羽
10	要紧	陶谦
11	转变观念	计策
12	谨记	空地
13	注意安全	点名
14	慎重	血书
15	分清	五万
16	防热	要不
17	少搞	长沙
18	两点论	体育场
19	认清形势	陈宫
20	要端正	不要

紧（肝心である）”

[+ 意志]

“分清主次（分をわきまえる）”,“牢记（しっかり心に刻む）”,“摆正位置（正しく位置付ける）”,“注意（用心する）”,“认清（見極める）”,“突出重点（中点を際立たせる）”,“做到（成し遂げる）”,“转变观念（観念を変える）”,“注意安全（安全に注意する）”,“慎重（慎重である / 慎重になる）”,“分清（はっきり区別する）”,“防热（熱を防ぐ）”,“少搞（やらない）”,“认清形势（形成を見極める）”,“谨记（しっかりと心に刻む）”

[その他]

“两点论（物事にはいい面と悪い面があるので、一面だけを見てはいけないという毛沢東の思想）”

“要”-“会”では、全体を通して訓戒を与えるような文脈で使われる語句が多い。[+ 義務] の語では、“必须”,“尤须”のように、義務性を表す語が上位に入っている。“要端正”に関しては形態素解析の間違いの可能性があり、本来的には“要”と“端正”の二語に分けるべきであるかもしれないが、いずれにせよ義務性の意味特徴を持つ語句が“要”-“会”で現れていることは間違いない。

[+ 意志] とした語句は、いずれも動作主が自らの意志で行う動作である。これらも使用される文脈を考えれば [+ 義務] との区別は極めて微妙であると言わざるを得ない。例えば“分清主次”という語はどのような文脈で使用されるかを考えてみれば、「私たちは事の軽重をわきまえるべきだ」「あなたは事の軽重をわきまえろ」というような、何らかの義務を負わせるような文脈になるだろう。即ち、[+ 意志] に分類した語であっても、本質的には [+ 義務] の文脈の中で使われている可能性が高いということである。その結果として、何らかの意志的な動作をさせようということだ。

[その他] には、リストアップされた中で唯一明確に名詞である“两点论”を分類した。“两点论”は毛沢東の思想の一部であり、「物事には良い面と悪い面が必ず備わっているから、人は一面的な見方をするのではなく、必ずその双方を見なければならない」とする考え方である。この語自体は体詞性であり、述詞性成分の多い他の語句とは異なっている。ただし、語の持つニュアンスは [+ 義務] と考えても差し支えない。このように考えると、苏（2017）モデルでの“要”-“会”は、[+ 義務] という意味特徴が色濃く反映された語が抽出されたとまとめることができるだろう。

一方で、王棟ら（2019）のモデルの結果はやや不可解である。小説データの中に含まれる、『三国演義』の二次創作小説に登場する地名や人名（例えば関羽など）が多くを占めてしまっているからである。

［人名］
“华雄”,“华佗”,“孔融”,“关羽”,“陶谦”,“陈宫”

［地名］
“长沙”

［一般語句］
“五万（五万）”,“要不（さもないと）”,“体育场（体育場）”,“郡（古い行政単位）”,“老将（老いた将軍）”,“太守（三国時代の地位の 1 つ）”,“血书（血判状）”,“扎寨（駐屯する）”,“斩（切断する、の古漢語）”,“计策（策略）”,“空地（空地）”,“点名（点呼）”,“不要（要らない、してはいけない）”

［一般語句］に分類した語でも、“老将”“扎寨”,“斩”などは三国演義系列の小説の語である可能性が非常に高い。

この結果については、単純にモデルの語数が少ないことから適切な結果が得られなかったと考えることもできる。ただし、これもまた“要”の何らかの性質を反映している可能性も否定できないことを、次節で論じる。

3.5. 結果の分析

3.5.1. 苏（2017）のモデル

苏（2017）のモデルで得られた結果に関しては、“会”+“要”,“会”-“要”,“要”-“会”いずれも先行研究にある程度符合すると解釈できるものとなっている。

“会”+“要”では、モダリティを表す副詞や助動詞、前後関係を表す副詞や記号については、“会”や“要”の性質を適切に反映していると言える。しかしながら、構造助詞の“的”や接続詞については、“会”や“要”の性質をどう反映しているのか不明瞭である。

“会”-“要”では、大別して［+ 長時間］［+ 推測］［- 意志］の意味特徴を持つ語句が出現した。［+ 長時間］に関して言えば、“要”には“将要（まもなく）”の意味があり、ある事象が短い時間のうちに起こることを表すことができる。一方で、“会”にはそのような制約はない。そのため、“会”の生起する文脈から“要”の生起する文脈を引いたとき、長い時間を表す語が残るのは自然なことだと言えるだろう。

［+ 推測］に関しては、先行研究に照らして考えればやや違和感が残る点である。“会”も“要”もいずれも可能性判断に使用することができ、推測モダリティを表す“可能”が“会”側の文脈でのみ生起するとは考えにくいからである。しかしながら、北京語言大学の BCC コーパス（<http://bcc.blcu.edu.cn/>2019/9/20 参照）で用例を検索すると、“可能会”が 3358 件ヒットするのに対して、“可能要”は 611 件と 1/5 の件数しかない。同様に、“也许会”は 2,783 件ヒットするのに対して、“也许要”は 383 件と 1/8 程度の件数しかない。つまり、仮に“会”と“要”双方に可能性判断のモダリティがあるとしても、モーダル副詞との共起頻度には大きな差がある可能性が高い。王牧（2018）においても、“会”の推測は“要”よりも論理性が高いのに対して、“要”は目の前の状況からの直感的な推測が多いことが指摘されており、そのような場合では推測モダリティを持つ語は使用されにくい可能性がある。word2vec の結果は、このような言語事実を反映しているのだろう。

［- 意志］の語が残るのは、“要”の持つ動作主の意志を表す意味特徴が取り除かれているからだと思われる。“要”には“我要这个（これをください）”や“我要喝水（水が飲みたい）”のように、何かを「必要とする」「やりたい」という意味がある。一方で、“会”では「なにかをしたい」という積極的な意志を表す用法は少なくとも中心的ではない[vii]。そのため、“会”から“要”を引いた場合には、「何らかの事象を積極的に行いたいという意志」を持たない語句が残ると考えられ、そ

の具体例として出現・消失を表す動詞や、マイナスのニュアンスを持つ心理活動動詞（フレーズ）が残ると考えられ、これも先行研究と合致している。

このように見ていくと、“会”-“要”の結果は言語学における先行研究とある程度合致しており、言語事実を反映していると言えるだろう。

“要”-“会”をした結果として、義務を表わし、訓戒を与えるような語が多く残ったことは、そのまま“要”の持つ義務モダリティを反映している。“要”-“会”を行った際に残ったのは、ほぼすべてが「義務」に関係する語であった。“必须”のように義務のモダリティを直接的に表す語に加えて、“分清主次”，“摆正位置”のように、譴責したり、訓導したりするような語が残されたことは特筆に値する。つまり、義務を課すような文脈で使われるような語句が“要”-“会”の結果になっているのである。“要”には「必要性」を表す基本用法がある一方、“会”には少なくとも中心的な用法としてはそのようなものがない。そのため、“要”から“会”を引いた場合には､義務性と関わる語が残ったのだろう。このように見ていくと、“要”-“会”は先行する言語学の研究と矛盾せず、両者をよく区別できている。

3.5.2. 王棟ら（2019）のモデル

王棟ら（2019）のモデルは、苏（2017）のモデルに比べると解釈が難しい結果が出ている。

“会”＋“要”では、代名詞が多く得られた点が苏（2017）のモデルとの最大の差異である。一見すると助動詞である“会”や“要”と代名詞はまったく性質が異なるように見えるが、純粋な語順だけを見れば分布は類似しており、首肯できる結果だとは言える。

助詞の“了”が“会”・“要”のベクトルの可算の結果として現れたこと、つまり、“会”の持つ意味特徴と“了”の持つ意味特徴が重なっていることには、大きく議論の余地がある。一般に可能性の“会”とアスペクト助詞の“了 1”と共起しづらいことが知られている。統語的にも“了”は動詞や文の後に置かれ､“会”は動詞の前に置かれるのが一般的であることから、両者を交換できる状況は少ないと思われるが、このような結果が出ている。“了 1”は完了や実現を表わす一方、可能性“会”はまだ未実現の事象を目的語に取ることが多く、両者の意味特徴が近いとは言いづらい。このモデルは、能力と可能性の“会”を区別しておらず、“了 1”と“了 2”の区別もしていない。そのため、極めて粗削りな結果しか得られておらず、文法について議論を深めるには不十分だと言える。

“会”-“要”や“要”-“会”では、一見して“会”や“要”の持つ機能と関係が見いだせない語が多く残っている。特筆すべきは“要”-“会”で、『三国演義』に由来する固有名詞や表現が非常に多く残されていることで、こちらも言語学の先行研究で論じられたことのないような傾向が出ていると言える。

なぜこのような結果が出ているのかについては、以下の二種類の理由が考えられる。第一に、やはりモデルの語数が少ないことである。22 億語という語数は、苏（2017）モデルにくらべても明らかに小さいものである。その結果、学習が不十分となり、加減算の結果にもノイズが混ざってしまったと考えられる。第二に、この結果は“会”や“要”の何らかの性質を反映している可能性がある。“要”-“会”をした結果､『三国演義』由来の語が大量に残るという現象について、例えば以下のような解釈が可能である。このモデルの元になった文章の中には、『三国演義』自体は入っていない。『三国演義』を下敷きにした二次創作小説が大量に収録されている。歴史を題材にした小説の二次創作であるからには、各登場人物の動向や末路には推測の余地はない。なぜなら、歴史的事実や一次創作の『三国演義』の記述によって、登場人物のたどる道筋は既に決まってしまっているからだ。換言すれば、人物の行動や結末については「推測」の余地がなく、いきおい作品は彼らの「意志」「思考」を中心に描写されることになる。つまり、『三国演義』の二次創作という作風全体に［- 推測］［+ 意志］の特徴があると考えることもできるのである。word2vec による実験は、この文体上の特徴から影響を受けて、このような結果を出し

た可能性がある。

とはいえ、モデル全体の語数の少なさの影響もまた否定できるものではない。また、「歴史小説の二次創作だから、推測が少なく意志が多い」という仮説も、全ての類似の文章に適用できるかどうかは未知数である。この点については、同種の文章を使った追調査が待たれる。

4. 結論

4.1. 結論

約 650 億語の現代語を使ったモデルである苏(2017)を使用した機能語同士の加減算では、先行する中国語研究とある程度矛盾しない結果が得られた。“会”+“要”では、モダリティを表す副詞や助動詞、前後関係を表す副詞や記号がリストアップされ、“会”や“要”の性質を適切に反映していると言える。“会”-“要”では［+ 長時間］［+ 推測］［- 意志］の意味特徴を持つ語がリストアップされ、“要”-“会”では［+ 義務］［+ 意志］の意味特徴を持ち、義務を課すような文脈で使われやすい語がリストアップされた。この結果からは、十分に“会”と“要”の持つ傾向をうかがい知ることができる。また、推測モダリティを持つ語との共起関係が“会”と“要”で異なっているという言語事実も、この結果から窺うことができるだろう。文法研究のヒントを得るという用途で word2vec を使用できる可能性は十分にある。

一方、王棟ら（2019）のモデルの結果は、言語学・中国語学の先行研究に合致するとは言い難い面もある。特筆するべきは“要”-“会”の結果であり、『三国演義』由来の語句が大量に得られた。これは、管見の限りでは、言語学・中国語学における“要”研究で言及されていない現象である。この結果に関しては、単純に学習語数が少ないことから深層学習が不十分であることが考えられる。しかしそれに加えて、『三国演義』の二次創作という作品形式に影響を受けているとも考えられ、文体と助動詞の使用の関係についての研究材料にもなり得るだろう。

4.2. 今後の課題

第一に、王棟ら（2019）モデルのように、個人で製作できる数十億語規模のモデルでは、あまり言語学の先行研究に合致した結果が出ないことである。つまり、個人レベルで作成できるモデルでは、機能語の特徴を学習しきれず、言語学の研究の実用に足るようなものにはならないおそれがある。

苏（2017）の約 650 億語のモデルでの計算結果は十分に言語学の先行研究と合致しており、コーパス言語学におけるコーパスのように、言語研究の支援ツールとして使用することは可能だと思われる。しかしながら、約 650 億語のデータを深層学習させるためには、専用のサーバーと高性能のコンピュータを長期間占有する必要があり、金銭的・物理的に個人の研究者が手軽に利用できるものではない。

実際の言語研究においては、苏（2017）モデルのような、既に作成済みの大規模なモデルをオンラインで入手し、使用することは決して難しくない。しかし、研究者自らがモデルを作成するにあたっては、ハード面の問題が避けられないだろう。

言語研究における深層学習の利活用においては、実用に足る大規模モデルの作成のために乗り越えるべきハード面の問題が残されている。この点についてどのような実践的な工夫ができるか探求することを今後の課題とする。

また、“了”のように、極めて多様な下位分類や機能を持つ語を、形態素解析の段階で「一語」として割り当ててしまっていることも、文法研究という目的からすると粗削りすぎると言える。文法議論のために word2vec を使用する場合、最低限“了 1”と“了 2”は別の語として登録するなど、前処理の段階で工夫しなければ、議論を深めることは難しい。例えば、「動詞」のタグを持つ語の直後に分布し、句点の直前に無い“了”を“了 1”、それ以外の“了”を“了 2”であると定義し、動詞の直後の“了”で文が終わっていれば“了 1 + 2”のタグをつけるなどして、前処理の段階から機械的に区別できるようにするなど、中国語学

の知見をモデルに反映させることで、文法研究にふさわしいモデルを作成することができる可能性が高くなるだろう。とはいえ、多機能な語は長く議論が続いているのが常であり、“了”の分布とその機能自体にも必ずしも共通の見解があるわけではない。そのため、タグ付け自体が安易には行えない側面がある。また、現実問題として機械学習では言語表面に現れる統語的な特徴以外を学習させることは困難であり、あまり複雑なタグをつけることは不可能である。どのようにタグをつけてデータを処理すれば、より言語の特徴を反映した結果を得ることができるのかを解明することが今後の最大の課題となるだろう。

word2vec をよりよいツールとして使うためには、文法研究に最適化したデータ処理の手法の開発が必要である。

参考文献

K. C. G. C. a. J. D. MikolovTomas. (2013).『Efficient estimation of word representations』. ICLR Workshop.
内田論 .（2018）.「word2vec による類義語抽出と FrameNet の比較：言語研究のための質的検証」.『言語統計を用いた認知言語学研究へのアプローチ』, 統計数理研究所 , pp. 41-51.
内田論 .（2019）.「単語分散表現におけるパラメーター変化の影響：word2vec を用いた事例研究」.『統計数理研究所共同研究リポート』413, pp. 31-42.
王其莉 .（2015）.「中国語の“会”に関する一考察―「Ⅰ . 能力」「Ⅱ . 長じる」ではない第Ⅲ類の“会”を中心に―. 判断のモダリティに関する日中対照研究」,『日中言語対照研究論集』17, pp. 217-239.
王棟 , 石田智裕 , 張婷 , 佐野洋 .（2019）.「単語分散表現の言語研究への利用　―中国語での事例報告―」.『電子情報通信学会技術研究報告』, pp. 31-36.
王牧 .（2018）.「現代中国語の " 要 " と " 会 " が表す認識モダリティの差異」.『言語情報科学』16, pp. 19-35.
斎藤康毅 .（2018）.『ゼロから作る Deep Learning ②　自然言語処理編』, オイラリー・ジャパン .
安本真弓 .（2019）.「中国語可能表現のメカニズム―“会”と“能”構文を中心に―」.『跡見学園女子大学文学部紀要』54, pp. 95-109.
王晓凌 .（2007）.「“会”与非现实性」.『语言教学与研究』第一期 , pp. 60-67.
郭继懋 , 郑天刚编 . (2002).『似同实异 : 汉语近义表达方式的认知语用分析』. 中国社会科学出版社 .
朱德熙 .（1982）.『语法讲义』. 商务印书馆 .
苏剑林 .（2017）.「不可思议的 word2vec」. <https://spaces.ac.cn/archives/4304>, 2019/8/28 閲覧 .
彭利贞 .（2007）.『现代汉语情态研究』. 中国社会科学出版社 .
吕叔湘編 .（1999）.『现代汉语八百词（增订本）』. 商务印书馆 .

データベース

「BCC 汉语语料库」. <http://bcc.blcu.edu.cn/>. 北京语言大学语言智能研究院 . 2019/9/20 閲覧 .

注

i　例えば、無料画像認識アプリケーション「google レンズ」は、一般のスマートフォンでも無料で問題なく使用できる。https://lens.google.com/

ii　google 社はニューラルネットワークを利用することで、自動翻訳の精度を大幅に上昇させた。word2vec の開発チームも当時 google に所属していた。
cf. https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

iii　word(target -2)、つまり、対象語の二語前に位置している語の意。w(t-1) なら対象語の一語前（直前）、w(t+1) なら対象語の一語後（直後）を指す。対象語の前後のどこまでを参照するかはパラメータであり、実験者によって決められる。

iv e.g.「AI 戦略 2019」日本政府の統合イノベーション戦略推進会議は、「人・産業・地域・政府全てに AI」と題し、2025 年までに AI 人材を 25 万人育成するとしている。
https://www.kantei.go.jp/jp/singi/tougou-innovation/dai4/sanko1.pdf

v 筆者は中国語非母語話者であるため、本稿の作例は、複数の中国語母語話者のネイティブチェックを受けている。

vi “逗号”。日本語の読点に近い。

vii “我会考虑的”のように、他者に対する承諾と言う形で意志を表す用法は指摘されている。cf. 彭 (2007 p. 142-143)。ただし、“要”の持つ積極的な意志とはニュアンスが異なるだろう。